Erasmus University Rotterdam

Erasmus School of Economics

BACHELOR THESIS

# Perfecting the Use of Clustering and Dimension Reduction

## A Comparison Study of Different Clustering and Dimension Reduction Strategies

*Author:*
J.C. Borges Soares (473966)

*Supervisor:*
C. Cavicchia
*Second assessor:*
M. van de Velden

July 5, 2020

**Abstract**

This research compares three different methods that cluster and reduce the number of dimensions of data. These methods are the tandem analysis, the factorial k-means analysis and the clustering and disjoint principal component analysis. The methods are compared by applying the analyses on simulated data and on micro-economic data. The simulated data set consists of two variables, masked by four normal drawn variables, where the goal is to retrieve the cluster structure set by the two original variables. The factorial k-means analysis best extracts the clustering structure and reduces the masking effect. The clustering and disjoint principal component analysis extracts the clustering structure, while having disjoint components, making the interpretation of the results less complicated. Applying the three methods on a micro-economic data set confirms this statement. When choosing between the factorial k-means analysis and the clustering and disjoint principal component analysis, one has to choose between having the better clustering or having results that are more convenient for interpretation.

# Contents

# 1 Introduction

The clustering of data and dimension reduction are two of the most commonly used analyses for exploring data. The first analysis focuses on partitioning big data sets into more easily interpretable subsets. Dimension reduction aims to extract the most relevant information of a data set, making the interpretation of the data set more efficient.

When dealing with data sets containing a large number of variables, interpreting and visualising the data becomes a complex task. In cases where a share of these variables is suspected to be irrelevant or seem to have a low contribution to the data structure, this becomes especially difficult. The multi-dimensional space also makes it more difficult to cluster the data. Clustering the data while also reducing the number of dimensions thus becomes a difficult task, however, this combination of reducing and clustering the data proves to be useful in several situations.

One way to cluster the data and reduce the number dimensions is called the *tandem analysis* (Arabie & Hubert, 1994). This analysis first uses *Principal Component Analysis (PCA)* to reduce the number of dimensions (Pearson, 1901). Sequentially the discrete *k-means* algorithm is applied to cluster the observations (Lloyd, 1982). De Soete & Carroll (1994) disapprove with this approach, due to the fact that the PCA can select certain dimensions that do not help extract the clustering structure of the data.

De Soete & Carroll (1994) introduce a new method to reduced the number of dimensions and extract the original clustering structure. The method they proposed is called the *reduced k-means* analysis. The first step in this analysis focuses on clustering the complete data set. This is done by creating centroids in a reduced subspace. The points in the full space are then added to the cluster corresponding to the centroid with the lowest euclidean distance to that point. Thereafter, the observations are projected into the lower dimension space. While De Soete & Carroll (1994) claim this method will prevent the original clustering structure from getting lost when reducing the number of dimensions, this, however, is not the case, as shown by Vichi & Kiers (2001).

Vichi & Kiers (2001) introduce a method called the *factorial k-means*. The factorial k-means analysis simultaneously clusters observations and reduces the number of dimensions. In contrast to the tandem analysis, which optimises two different objective functions (which can contradict each other), the factorial k-means uses one objective function to simultaneously recover the factor loadings of the different dimensions and cluster the observations.

This method is expanded on by Vichi & Saporta (2009). They introduce the *Clustering and Disjoint Principal Component Analysis (CDPCA)*, which simultaneously reduces the number of dimensions and clusters the observations, similar to the factorial k-means. The difference being that this method clusters the different variables used. This method can especially be helpful when trying to interpret data sets with both a large number of observations and a large number of variables.

Many clustering and dimension reduction methods have been introduced through the years. In this research I will show which method gives the best results when applied on simulated data and

real economic data. The central research question answered in this research therefore is: "How can observations and variables, from simulated and economic data, best be clustered while also reducing the number of dimensions?"

The remainder of this research is structured as follows. In Section 2, I discuss the methods used to cluster and reduce the number of dimensions of the data. Section 3 compares the performance of the three different methods when applied on simulated (masked) data. Afterwards, I compare the methods on a more realistic data set containing different micro- economic variables in Section 4. I conclude this research with remarks, limitations and possibilities regarding improvement of the research in Section 5.

# 2 Methodology

In this section, I discuss three methods to cluster data observations, while also reducing the dimension space. The first method, the *tandem analysis*, first reduces the dimensions using PCA (Pearson, 1901). Subsequently, the k-means algorithm is used to cluster the data in this reduced space (Lloyd, 1982). The remaining two methods, *factorial k-means* and *CDPCA*, simultaneously reduce the dimensions and cluster the data.

## 2.1 Factorial k-means

The factorial k-means, first introduced by Vichi & Kiers (2001), simultaneously reduces the number of dimensions and clusters the observations. The number of clusters and number of dimensions have to be specified in advance. The factorial k-means method minimises the within cluster deviance. This results in the model given in Equation (1).

$$\mathbf{XAA}' = \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}' + \mathbf{E} \tag{1}$$

Here $\mathbf{X}$ is the $(n \times k)$ data matrix, where $n$ are the number of observations and $k$ are the number of variables. $\mathbf{U}$ is an $(n \times c)$ binary matrix, where $c$ are the, in advance specified, number of clusters. If $u_{ij}$ gives the element on the $i$'th row and the $j$'th column of matrix $U$, then $u_{ij} = 1$ if observation $i$ belongs to cluster $j$, and $u_{ij} = 0$ otherwise. $\mathbf{A}$ is a $(k \times m)$ column wise orthonormal matrix, were $m$ are the number of dimensions describing the clustering. $\mathbf{A}$ gives the coefficients of the linear combinations of the different variables, the factor loadings. $\mathbf{Y}$ gives the objective scores of all the different observations. The centroids of all the clusters are given by $\bar{\mathbf{Y}}$. Finally, $\mathbf{E}$ is the matrix of error components.

Equation (1) shows an orthogonal projection onto a subspace spanned by the columns of $\mathbf{A}$. The locations of the projections are given by the columns of $\mathbf{Y} = \mathbf{XA}$. With these components, we try to find a clustering that minimises the distance between the points in a cluster and the centroid of the corresponding cluster (within cluster deviance). Minimising the distance is done by minimising

the error components. Minimising the error components is achieved by minimising the objective function

$$\mathrm{F}(\mathbf{A}, \mathbf{U}, \bar{\mathbf{Y}}) = \|\mathbf{XAA}' - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}'\|^2 = \|\mathbf{XA} - \mathbf{U}\bar{\mathbf{Y}}\|^2, \tag{2}$$

subject to the two constraint that $\mathbf{U}$ is binary and has exactly one unit element per row and that $\mathbf{A}'\mathbf{A} = \mathbf{I}_m$. Here $\mathbf{I}_m$ is an $(m \times m)$ identity matrix. The optimal $\bar{\mathbf{Y}}$ can be written in terms of $\mathbf{A}$ and $\mathbf{U}$ as follows,

$$\bar{\mathbf{Y}} = (\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{XA}. \tag{3}$$

Substituting $\bar{\mathbf{Y}}$ into the objective function stated in Equation (2), a objective function depending solely on $\mathbf{A}$ and $\mathbf{U}$ is retrieved, as stated in Equation (4).

$$\mathrm{F}(\mathbf{A}, \mathbf{U}) = \|\mathbf{XA} - \mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{XA}\|^2 \tag{4}$$

Now we use the fact that the squared norm can be rewritten using the trace and transpose operations, $\|\mathbf{X}\|^2 = \mathrm{tr}(\mathbf{X}'\mathbf{X})$ (Hastie et al., 2001). The objective function can be written as the following function:

$$\mathrm{F}(\mathbf{A}, \mathbf{U}) = \mathrm{tr}(\mathbf{A}'\mathbf{X}'\mathbf{XA}) - \mathrm{tr}(\mathbf{A}'\mathbf{X}'\mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{XA}). \tag{5}$$

The objective function now consists of two separate parts. The first part is equal to the total deviance of $\mathbf{XA}$. The second part equals the between cluster deviance.

To minimise the objective function stated in Equation (2) (equivalently (5)), standard OLS cannot be used. We therefore use the alternating least squares (ALS) algorithm (Vichi & Kiers, 2001). The ALS algorithm is described in the following steps:

**Initial Step (Step 0):** In this step, the values of $\mathbf{A}$, $\mathbf{U}$ and $\bar{\mathbf{Y}}$ are initialised. The values are chosen randomly, however, the values for $\mathbf{A}$ and $\mathbf{U}$ have to meet the constraints.

**Step 1:** Given the values of $\mathbf{A}$ and $\bar{\mathbf{Y}}$, the objective function of Equation (2) is minimised with respect to $\mathbf{U}$. This is done for the rows of $\mathbf{U}$ independently. The column of $\mathbf{U}$ is set to 1 when it gives the lowest value for our objective function compared to setting the other columns to 1. This can be mathematically written as: $\mathrm{F}(\mathbf{A}, [u_{ij}]) = \min\{\mathrm{F}(\mathbf{A}, [u_{iv}])\} \; \forall \; v = 1, ..., c$, where $[u_{ij}]$ equals the element on the $i$'th row and the $j$'th column of $\mathbf{U}$.

**Step 2:** In this step $\mathbf{A}$ and $\bar{\mathbf{Y}}$ will be updated for a given $\mathbf{U}$. Instead of using Equation (2), Equation (5) is used. $\mathbf{A}$ will be optimised by taking the first $m$ eigenvectors of $\mathbf{X}'(\mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}' - \mathbf{I}_n)\mathbf{X}$, where $\mathbf{I}_n$ is the identity matrix of the $n$'th order. When the optimal $\mathbf{A}$ is found, the optimal $\bar{\mathbf{Y}}$ can be calculated by inserting the optimal $\mathbf{U}$ from **Step 1** and the optimal $\mathbf{A}$ in Equation (3).

**Step 3:** After completing the calculation of the optimal $Y$, the objective value using the calculated values of $\mathbf{A}$ and $\mathbf{U}$ is computed. The new objective value will be decreased compared to the old

objective value. **Step 1** and **Step 2** are being repeated until the objective value does not decrease by more than $10^{-5}$.

The algorithm, described in the steps above, always decreases the objective function. A problem that may occur is that the minimising algorithm gets caught in a local minimum. Ten Berge (1993) describes two possibilities to counter this problem. The first option is to initiate our starting values of $\mathbf{A}$, $\mathbf{U}$ and $\bar{\mathbf{Y}}$, closer to the global minimum. Due to the algorithm always decreasing the objective value, the algorithm cannot get caught in a local minimum when our initial values result in a objective value lower than the local minima. The second option is to use more (random) starting points. Using several initial values yields in a higher probability of finding the global minimum. The last option will be implemented in this research.

## 2.2 Clustering and disjoint principal component analysis

The factorial k-means clusters observations while simultaneously reducing the number of dimensions, leaving us with a reduced space which is expressed by a combination of variables. These loadings can overlap, meaning that one variable can explain more than one dimension, which is hard to interpret. The interpretation is especially complex when there are a large number of variables. To prevent the interpretation becoming overly complex, the following method, first introduced by Vichi & Saporta (2009), is utilised. The method is called the *Clustering and Disjoint Principal Component Analysis (CDPCA)*. The CDPCA is similar to the factorial k-means in the fact that it clusters observations while simultaneously reducing the number of dimensions, however, the CDPCA also clusters the explaining variables. This means that the CDPCA averts the problem of having overlapping component loadings, which makes the results easier to interpret.

The model is defined in Equation (6). Where the factorial k-means analysis minimises the within-in cluster deviance, the CDPCA maximises the between-cluster deviance. The parameters stated are equivalent to the ones stated in Section 2.1.

$$\mathbf{X} = \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}' + \mathbf{E} \tag{6}$$

However, the constraints for the parameter $\mathbf{A}$ are different in this case. The CDPCA uses the column wise orthonormal matrix $\mathbf{A}$, equivalent to the factorial k-means model. The difference is that the matrix satisfies two constraints: each row has exactly one nonzero value and the squared sum of each column equals 1. Through the first constraint, we observe that a variable can only explain one component, thus clustering the variables also.

To estimate the parameters stated in Equation (6) the least-squares objective function,

$$\mathrm{F}(\mathbf{U}, \bar{\mathbf{Y}}, \mathbf{A}) = \|\mathbf{X} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}'\|^2, \tag{7}$$

is minimised, subject to $\mathbf{U}$ being binary and row stochastic and $\mathbf{A}$ following the two constraints

stated above.

Observing Equation (7), the decomposition of the statement can be done as follows:

$$\|\mathbf{X} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}'\|^2 = \|\mathbf{X}\|^2 - \|\mathbf{U}\bar{\mathbf{Y}}\mathbf{A}'\|^2. \tag{8}$$

The mathematical proof of this decomposition is stated below. We here use that $\bar{\mathbf{Y}} = \bar{\mathbf{X}}\mathbf{A}$.

$$\begin{aligned}
\|\mathbf{X} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}'\|^2 + \|\mathbf{U}\bar{\mathbf{Y}}\mathbf{A}'\|^2 &= \mathrm{tr}\{\Big[\mathbf{X} - \mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}'][\mathbf{X} - \mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}']'\Big\} + \mathrm{tr}\{\Big[\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}'][\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}']'\Big\} \\
&= \mathrm{tr}\{\mathbf{X}\mathbf{X}'\} - 2\mathrm{tr}\{\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}'\mathbf{X}'\} + 2\mathrm{tr}\{\mathbf{U}'\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}'\bar{\mathbf{X}}'\} \\
&= \mathrm{tr}\{\mathbf{X}\mathbf{X}'\} - 2\mathrm{tr}\{\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}'\mathbf{X}'\} + 2\mathrm{tr}\{\mathbf{U}'\mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{X}\mathbf{A}\mathbf{A}'\bar{\mathbf{X}}'\} \\
&= \mathrm{tr}\{\mathbf{X}\mathbf{X}'\} = \|\mathbf{X}\|^2
\end{aligned} \tag{9}$$

Using the decomposition proved in Equation (9). We can, instead of minimising Equation (7), maximise the term

$$\|\mathbf{U}\bar{\mathbf{Y}}\mathbf{A}'\|^2. \tag{10}$$

This term equals the between-cluster deviance of $\mathbf{X}$ clustered by $\mathbf{U}$. The final simplification is using $\mathbf{Y} = \mathbf{X}\mathbf{A}$ to rewrite Equation (10).

$$\|\mathbf{U}\bar{\mathbf{Y}}\mathbf{A}'\|^2 = \|\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}'\|^2 = \mathrm{tr}\{\Big[\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}'][\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}']'\Big\} = \mathrm{tr}\{\big[\mathbf{U}\bar{\mathbf{X}}\mathbf{A}][\mathbf{U}\bar{\mathbf{X}}\mathbf{A}]'\big\} = \|\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\|^2 \tag{11}$$

Using this result, we thus solve the problem by maximising the between-cluster deviance $\|\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\|^2$, subject to the two constraints of $\mathbf{U}$ being binary and row stochastic and $\mathbf{A}$ being column wise orthonormal and having exactly one non-zero element per row. This is done using an ALS algorithm, similar to the factorial k-means algorithm.

Comparing the factorial k-means analysis and the CDPCA, we see that the two analyses use quite similar methods of clustering. However, where the factorial k-means minimises the within-cluster deviance of $\mathbf{Y}$, the CDPCA maximises the between-cluster deviance. The reason the between-cluster deviance is used in the CDPCA is due to the CDPCA wanting to define factors of maximal variance to cluster the variables. This is guaranteed if the between-cluster deviance is maximised.

Moreover, the CDPCA distincts itself from the factorial k-means in the fact that the CDPCA also clusters the variables, which can be informative when operating large data sets. The interpretation is less complicated when the factor loadings of the PCA are disjoint.

### 2.3 Choosing the number of dimensions

When using the factorial k-means analysis and CDPCA, the number of dimensions and the number of clusters have to be specified in advance. Assuming the number of clusters are constant and specified in advance, we can determine the number of dimensions by observing the eigenvalues and explained variance of each component when applying PCA. When applying PCA on the complete

data set, objective scores of each observation are determined for each dimension. The objective scores can be used to derive the eigenvalues of the principal components. The eigenvalues are calculated by computing the the covariance matrix of the objective scores (Heij et al., 2004). Thereafter, the eigenvalues are calculated by solving Equation 12 to $\lambda$ (Poole, 2006).

$$\det(\mathbf{\Sigma} - \lambda \mathbf{I}_k) = 0 \tag{12}$$

Here $\mathbf{\Sigma}$ is the covariance matrix of the PCA objective scores and $\mathbf{I}_k$ is a $k \times k$ identity matrix.

The eigenvalues are used to calculate the percentage of variance explained by each principal components. This is done dividing each eigenvalue by the total sum of the components' eigenvalues. The cumulative sum of the explained variance can be calculated by adding the antecedent percentages and adding them to the percentage of the component that is being calculated. The criterion used to choose the number of components is based on this cumulative percentage of variance explained. I choose the number of dimensions by having the components explain at least 80% of the total variance. That is to say that the cumulative sum of explained variance has to be larger than 80%.

# 3 Masking of the clustering structure using simulated data

In this section I use simulated data to compare the performance of the three methods. The section is split up into three sections, the tandem analysis, factorial k-means analysis and CDPCA are all applied on the simulated data.

The simulated data set contains forty two observations and six variables. The first two variables contain the location of the observations, as can be seen in Figure 1 (page 8). The locations are chosen to represent a well-defined clustering structure with three separate clusters (Gordon, 1999). The centroids of the three clusters are located on the vertices of an equilateral triangle with sides of length six. The remaining four variables are drawn from a normal distribution with mean zero and a standard deviation of six.

The three analyses are conducted in R, using the *clustrd* package by Markos et al. (2019) and the *biplotbootGUI* package by Librero et al. (2019). The R scripts can be found in Appendix A. The used data can be found in Appendix B.

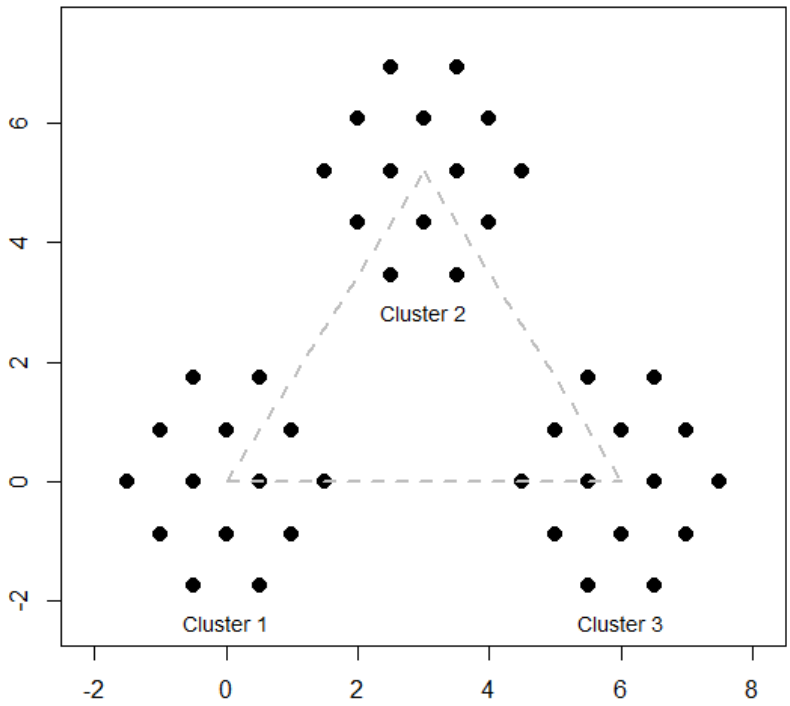


Figure 1: Plot of the first two variables

### 3.1 Tandem analysis

To mask the clustering structure, four normal drawn variables are added to the two existing variables. These variables are irrelevant and thus will help to describe the problem I am trying to solve. Observing Figure 2 (page 9), we can see that applying the k-means algorithm on all observations, including the observations retrieved from the four irrelevant variables, results in a non-optimal clustering of the observations. The clustering structure set by the two-dimensional variables is masked by the six-dimensional data set. To be more precise, the k-means algorithm only correctly clusters twenty of the forty two observations, as seen in Figure 2.

To avert the masking effect, one would think using PCA on the data set would reduce the masking effect. This because the PCA retrieves the most relevant information from the data set. The results of applying PCA on the six dimensional data set and then applying k-means to the first two, three, four and five components are given in Figure 3 (page 10). The plots with three or more components are plotted on the two original variables. Figure 3 shows that using PCA to extract the relevant information does not help solve the problem, but increasing the number of used components does increase the ability to retrieve the original cluster structure. Figure 3.a shows that

the first two components do not show the the three well-separated classes described by the first two variables. This means that the PCA fails to retrieve the most relevant information. The clustering in this figure also does not correctly cluster the observations.

When looking at Figure 3.b-3.d, we observe that increasing the number of components used, increases the performance of the clustering method. Figure 3.b shows that 14 of the 42 observations are incorrectly clustered. This number decreases to 11 when taking four components into account. When using five components, as seen in Figure 3.d, 8 of the 42 observations are incorrectly clustered.

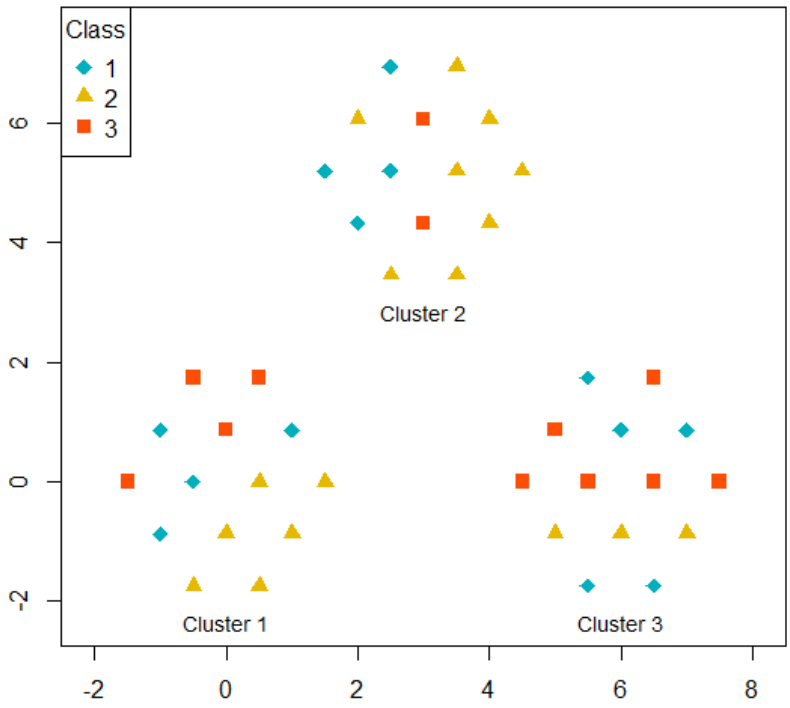


Figure 2: K-means algorithm applied on the full simulated data set

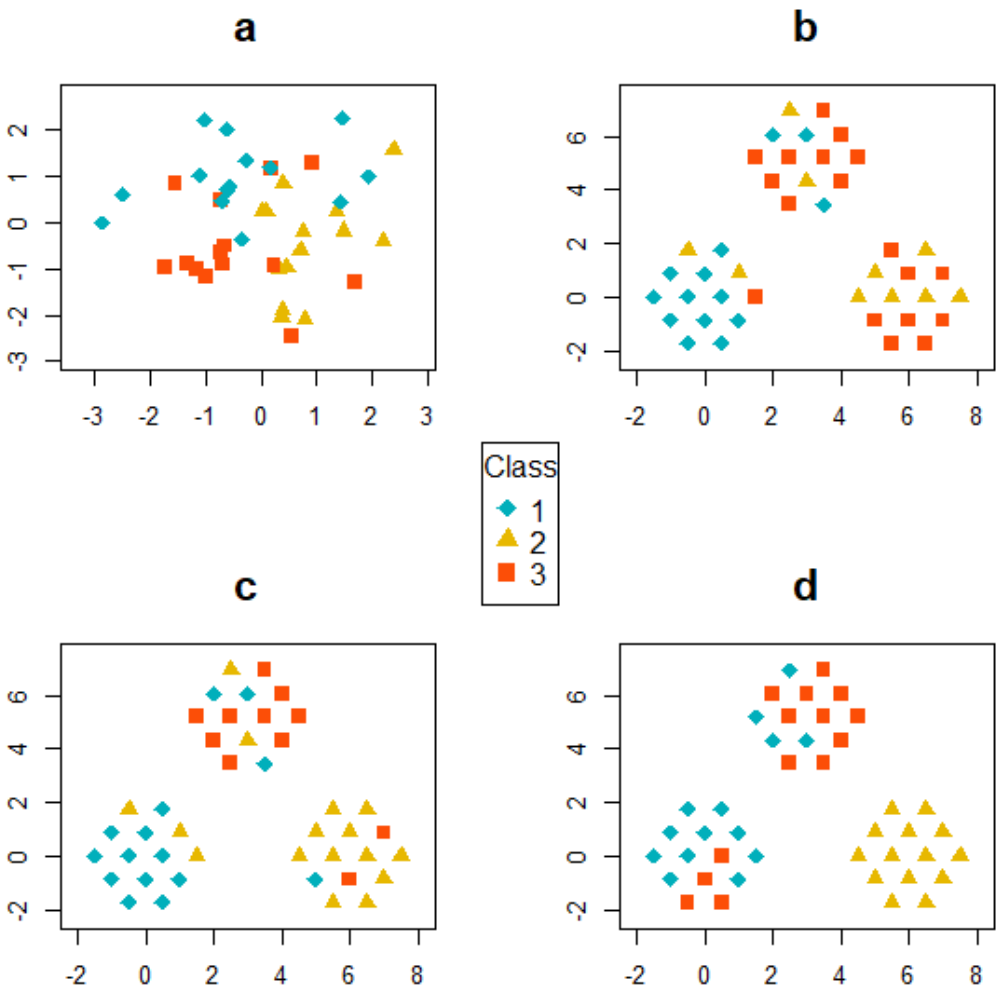


Figure 3: Tandem analysis applied on the first a) two principal components b) three principal components c) four principal components d) five principal components

Table 1 shows the eigenvalues, explained total variance and cumulative variance of the six principal components. The explained cumulative percentage of the total variance is less than 50 percent for the first two components. The first four components all explain more than 15 percent of the total variance. This justifies the results shown in Figure 3. Applying the k-means algorithm on more components significantly enhances the ability to retrieve the original cluster structure.

Table 1: Eigenvalue, explained total variance and cumulative variance of the six principal components

| Components | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Eigenvalue | 1.438 | 1.398 | 1.102 | 1.021 | 0.663 | 0.380 |
| % Variance | 23.961 | 23.298 | 18.361 | 17.009 | 11.046 | 6.325 |
| % Cumulated | 23.961 | 47.259 | 65.620 | 82.629 | 93.675 | 100 |

### 3.2 Factorial k-means

The factorial k-means model, introduced by Vichi & Kiers (2001), is applied on the simulated data. The number of clusters are set to three and the number of dimensions to two. The ALS algorithm ran for 100 different starting values, as suggested by Vichi & Kiers (2001).

Observing Figure 4, we see that the factorial k-means near perfectly captures the cluster structure stated by the two relevant variables. Not only does the method cluster all forty two observations correctly, but the separated clustering structure is also captured by this method.

This result is also shown in Table 2 (page 12), where the factor loadings (correlation between dimensions and variables) are shown. We here see that the first dimension is highly correlated with the first variable and the second dimension is highly correlated with the second variable. The four added 'mask' variables all have a low correlation with the dimensions.

Using the simulated data, the difference between the tandem analysis and factorial k-means is shown clearly. We see that when applying the tandem analysis, the masking of the clustering structure is not reduced. The factorial k-means method does reduce the masking effect, if not completely eliminates it.

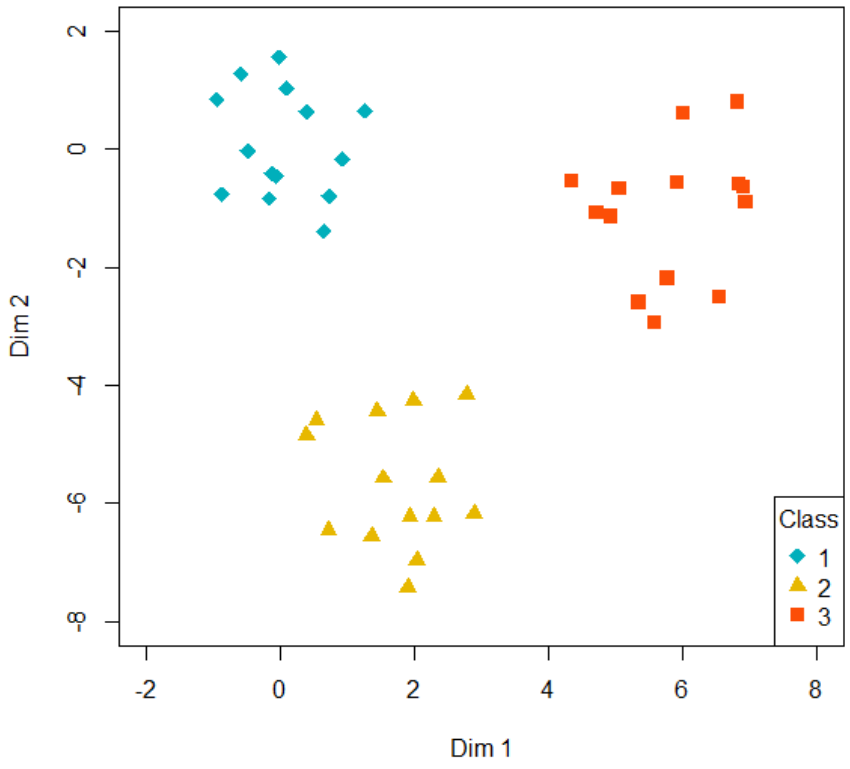


Figure 4: Clustering of the simulated data using factorial k-means

Table 2: Factor loadings of factorial k-means applied on simulated data (factor loadings bigger than 0.3 in absolute value are highlighted)

| | V1 | V2 | V3 | V4 | V5 | V6 |
|---|---|---|---|---|---|---|
| Dim 1 | **0.977** | -0.203 | -0.040 | -0.020 | -0.050 | -0.021 |
| Dim 2 | -0.205 | **-0.977** | -0.058 | 0.022 | 0.013 | -0.006 |

### 3.3 Clustering and Disjoint PCA

Observing the results of the factorial k-means analysis, especially Table 2, we see that the different variables are correlated with both dimensions. This can make the plot in Figure 4 hard to interpret. To make this interpretation easier, we aim to create disjoint dimensions, which brings us to the Clustering and Disjoint PCA (Vichi & Saporta, 2009). The ALS algorithm of the CDPCA analysis ran for 30 different starting values, as suggested by Vichi & Saporta (2009).

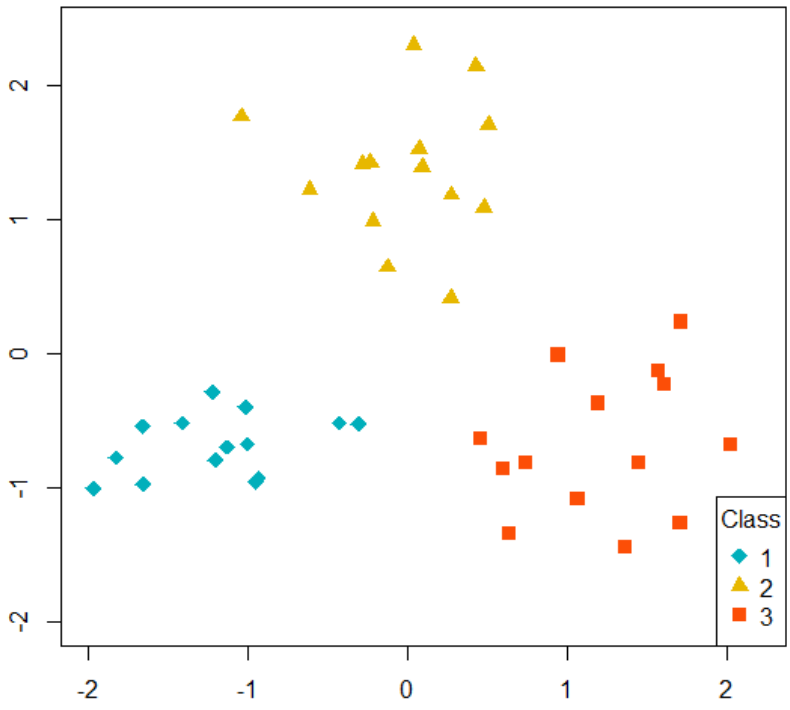


Figure 5: Clustering of the simulated data using CDPCA

Table 3: Factor loadings of CDPCA applied on simulated data (factor loadings bigger than 0.3 in absolute value are highlighted)

| | V1 | V2 | V3 | V4 | V5 | V6 |
|---|---|---|---|---|---|---|
| Dim 1 | **0.960** | 0 | 0 | -0.119 | 0.253 | 0 |
| Dim 2 | 0 | **0.946** | 0.138 | 0 | 0 | 0.293 |

Observing Figure 5 (page 12), we see that the CDPCA also succeeds in retrieving the clustering structure from the masked data. However, the clustering structure is not as clear as that of the original two variables or the factorial k-means analysis. This can also be seen when observing Table 3 (page 12). We first notice that the variables are only correlated with one dimension, which is implied by the restriction on the factor loading matrix $\mathbf{A}$. Furthermore, Table 3 shows that the loadings of the four normal variables are higher than in the factorial k-means analysis. This is not unexpected, because the CDPCA is more restrictive than the factorial k-means.

Using the simulated data, the difference between the methods is shown clearly. Based on the results recieved when applying the tandem analysis, it can be concluded that the masking of the clustering structure is not reduced. The factorial k-means method does reduce the masking effect, if not completely eliminates it. The CDPCA reduces the masking effect, however, the reducing of the masking effect is not as efficient as using the factorial k-means analysis.

# 4 Application on Economic Data

Following the researches of Vichi & Kiers (2001) and Vichi & Saporta (2009), I will apply the methods on economic data to test the clustering ability of the three methods. A data set containing micro-economic data has been considered.

The data set used, consists of micro-economic performances of 77 different economies. These economies are described using twelve indicators. The countries used in the analysis are shown in Table 4 and the used variables used are shown in Table 5 (page 14).

The variables consists of indices, which are collected by Numbeo (2020). These indices are relative values and are compared to values of New York City. Taking *Crime Index* as example, this denotes that a index value larger than 100 corresponds to a higher crime rate compared to New York City. The same reasoning is applied when observing indices lower than 100.

The countries will be clustered into seven clusters, as done by Vichi & Saporta (2009), who used a data set of similar size. When applying the three analyses, the data is zero centred and scaled (Markos et al., 2019). The zero centring subtracts the mean from all the objective scores to move them around the 0 coordinate. The scaling makes it that each variable is standardised such that the variance equals 1.

The different analyses are conducted in R, using the *clustrd* package by Markos et al. (2019) and the *biplotbootGUI* package by Librero et al. (2019). The codes can be found in Appendix C and the used data can be found in Appendix D.

To determine the number of dimensions, we applied PCA on the data set. The eigenvalues, explained variance and cumulative explained variance are shown in Table 6 (page 14). This table shows that, to meet the criterion stated in Section 2.3, four components are chosen to represent our data. Figure 6-8 all plot the first two components, because these explain the most variance.

Table 4: Alphabetically ordered list of countries and corresponding labels of the economic data set

| |
|---|
| Argentina (ARG), Australia (AUS), Austria (AUT), |
| Bangladesh (BAN), Belarus (BEL), Belgium (BEG), Brazil (BRA), Bulgaria (BUL), |
| Canada (CAN), Chile (CHI), China (CHN), Colombia (COL), Croatia (CRO), Cyprus (CYP), Czech Republic (CZE), |
| Denmark (DEN), |
| Ecuador (ECU), Egypt (EGY), Estonia (EST), |
| Finland (FIN), France (FRA), |
| Georgia (GEO), Germany (GER), Greece (GRE), |
| Hong Kong (HON), Hungary (HUN), |
| Iceland (ICE), India (IND), Indonesia (INO), Iran (IRA), Ireland (IRE), Israel (ISR), Italy (ITA), |
| Japan (JAP), Jordan (JOR), |
| Kazakhstan (KAZ), Kenya (KEN), Kuwait (KUW), |
| Latvia (LAT), Lebanon (LEB), Lithuania (LIT), |
| Malaysia (MAL), Mexico (MEX), Morocco (MOR), |
| Netherlands (NET), New Zealand (NZE), North Macedonia (NMA), Norway (NOW), |
| Oman (OMA), |
| Pakistan (PAK), Panama (PAN), Peru (PER), Philippines (PHI), Poland (POL), Portugal (POR), |
| Qatar (QAT) |
| Romania (ROM), Russia (RUS), |
| Saudi Arabia (SAU), Serbia (SER), Singapore (SIN), Slovakia (SLA), Slovenia (SLE), South Africa (SAF), South Korea (SKO), |
| Spain (SPA), Sri Lanka (SRI), Sweden (SWE), Switzerland (SWI), |
| Taiwan (TAI), Thailand (THA), Turkey (TUR), |
| Ukraine (UKR), United Arab Emirates (UAE), United Kingdom (UK), United States (US), |
| Vietnam (VIE). |

Table 5: Variables of the economic data set

| |
|---|
| Ages 0 to 14 years (%) |
| Over the age of 65 (%) |
| Crime Index |
| Cost of Living Index |
| Rent Index |
| Groceries Index |
| Restaurants Index |
| Health Care Index |
| Quality of Life Index |
| Purchasing Power Index |
| Traffic Index |
| Pollution Index |

Table 6: Eigenvalue, explained total variance and cumulative variance of the twelve principal components derived from the economic data

| Components | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Eigenvalue | 7.136 | 1.608 | 0.786 | 0.749 | 0.651 | 0.350 | 0.279 | 0.190 | 0.127 | 0.092 | 0.022 | 0.010 |
| % Variance | 59.468 | 13.402 | 6.548 | 6.243 | 5.428 | 2.915 | 2.325 | 1.586 | 1.054 | 0.765 | 0.184 | 0.082 |
| % Cumulated | 59.468 | 72.869 | 79.417 | 85.660 | 91.088 | 94.004 | 96.329 | 97.915 | 98.969 | 99.734 | 99.918 | 100 |

## 4.1 Tandem analysis

Applying the tandem analysis on the economic data set, we first apply PCA on the 12-dimensional data. I choose to use the first four principal components, corresponding to the components that cumulatively explain 85.66% of the variance. Thereafter, we cluster the 77 countries into seven clusters. This is done by applying k-means on the PCA objective scores. The factor loadings of the first four components are shown in Table 7.a (page 16). Figure 6 shows a plot of the first two components.

Observing the factor loadings, we can see that the first component is highly positively correlated with the cost of living, the cost of groceries, the restaurant prices, the quality of life and the purchasing power. These variables are all indicators for the prices within a country. The amount of pollution has a large negative correlation with the first component. The percentage of children and adolescents, cost of living, rent price, price of groceries and the traffic commute time all have a large positive correlation with the second component. The percentage of elderly is highly negatively correlated with the second component. We can see that the second component is not as focused on the price of products, but focuses more on the demography and infrastructure of an country. We also see that the grocery prices and cost of living characterises both components. The third component is highly positively correlated with the percentage of elderly, the crime rate and the health care quality. The rent price and the amount of pollution have a large negative correlation with this component. The percentage of children and adolescents has a large positive correlation with the fourth component, while the health care quality and traffic commute time are highly negatively correlated with this component.

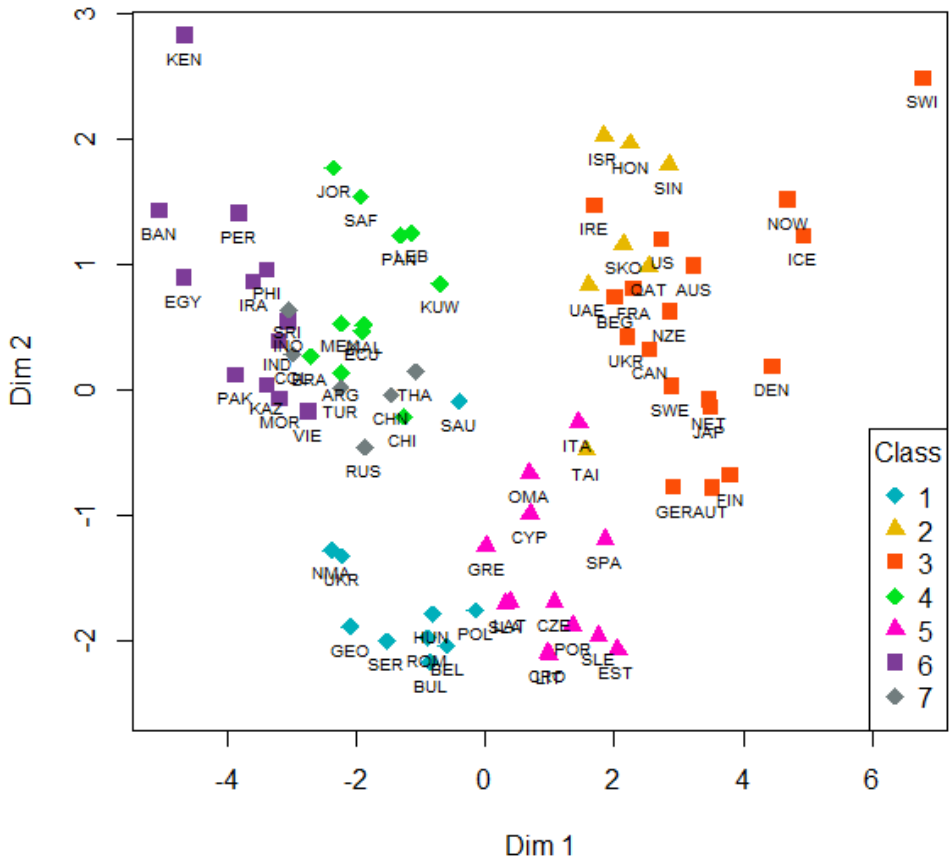


Figure 6: First two components of the tandem analysis applied on the economic data set

Table 7: Factor loadings of the four components ($c = 4$) retrieved from applying the a) tandem analysis b) factorial k-means analysis c) CDPCA on the economic data set (factor loadings bigger than 0.3 in absolute value are highlighted)

| | a) Tandem | | | | b) Factorial k-means | | | | c) CDPCA | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Dim 1 | Dim 2 | Dim 3 | Dim 4 | Dim 1 | Dim 2 | Dim 3 | Dim 4 | Dim 1 | Dim 2 | Dim 3 | Dim 4 |
| Ages 0 to 14 years (%) | -0.2594 | **0.3891** | 0.1664 | **0.3740** | 0.1173 | 0.2870 | **-0.3930** | 0.1267 | 0 | **0.5563** | 0 | 0 |
| Over the age of 65 (%) | 0.2397 | **-0.3868** | **0.3663** | -0.2672 | 0.0725 | **0.3134** | **-0.3601** | 0.2877 | 0 | **-0.4767** | 0 | 0 |
| Crime Index | -0.2308 | 0.2737 | **0.6244** | 0.2449 | -0.0860 | -0.0167 | 0.1755 | -0.0577 | 0 | **0.4850** | 0 | 0 |
| Cost of Living Index | **0.3345** | **0.3057** | -0.0229 | -0.0019 | **0.7270** | 0.1013 | **0.3934** | -0.0815 | **0.5302** | 0 | 0 | 0 |
| Rent Index | 0.2761 | **0.3418** | **-0.3571** | -0.0783 | -0.1171 | 0.1999 | -0.0719 | 0.0760 | **0.4789** | 0 | 0 | 0 |
| Groceries Index | **0.3057** | **0.3528** | -0.0347 | -0.1815 | **-0.3623** | -0.0790 | -0.2938 | **-0.5561** | **0.5075** | 0 | 0 | 0 |
| Restaurants Index | **0.3222** | 0.2402 | 0.0471 | 0.2790 | **-0.3699** | **-0.2994** | 0.2051 | **0.6055** | **0.4817** | 0 | 0 | 0 |
| Health Care Index | 0.2399 | 0.1132 | **0.3797** | **-0.4947** | -0.0208 | -0.1184 | 0.1171 | 0.2417 | 0 | 0 | 0 | **0.6170** |
| Quality of Life Index | **0.3425** | -0.1588 | 0.1554 | 0.2380 | **-0.3044** | **0.6528** | **0.3485** | -0.1835 | 0 | 0 | **0.7353** | 0 |
| Purchasing Power Index | **0.3246** | 0.1457 | 0.0023 | 0.1615 | 0.2351 | 0.0183 | **-0.4926** | 0.1377 | 0 | 0 | 0 | **0.7870** |
| Traffic Index | -0.2432 | **0.3788** | 0.1951 | **-0.4776** | -0.0577 | 0.0745 | 0.1247 | -0.0079 | 0 | **0.4776** | 0 | 0 |
| Pollution Index | **-0.3117** | 0.1700 | **-0.3323** | -0.2374 | -0.1076 | **0.4764** | 0.0576 | **0.3136** | 0 | 0 | **-0.6777** | 0 |

## 4.2 Factorial k-means

When using simulated data, the factorial k-means method showed its strength in retrieving the original clustering structure from masked data. I used 100 random draws to initiate our parameters, as suggested by Vichi & Kiers (2001). The result when applying the method on economic data is plotted in Figure 7 (page 17). The factor loadings are shown in Table 7.b.

Observing the factor loadings, we see that the first dimension is largely positively correlated with the cost of living and largely negatively correlated with the price of groceries, the restaurant prices and the quality of life. The second dimension is largely positively correlated with percentage of elderly people, the quality of life and the amount of pollution. This dimension is largely negatively correlated with the restaurant price. The third dimension is highly positively correlated with the cost of living and the quality of life and has a large negative correlation with both the percentage of children and adolescents and the percentage of elderly. The purchasing power has a large negative correlation with this dimension as well. The fourth dimension is highly positively correlated with the restaurant prices and the amount of pollution. This dimension has a large negative correlation with the grocery prices.

## 4.3 Clustering and Disjoint PCA

Lastly, the CDPCA was applied on the economic data set. The simulated data showed that this method does succeed in retrieving the masked cluster structure, and has a very nice advantage, in its ability to also cluster the variables, making the interpretation of the results less complex. I used 30 random draws to initiate our parameters, as suggested by Vichi & Saporta (2009).

When looking at the factor loadings of the first two components in Table 7.c, we can clearly see the disjoint dimensions. All the variables used have a large correlation with their corresponding dimension. The first component is positively correlated with the cost of living, rent pricing, price of groceries and the price at restaurants. Equal to the tandem analysis, this component is more fixated on the variables describing prices. There are no variables having a negative correlation with

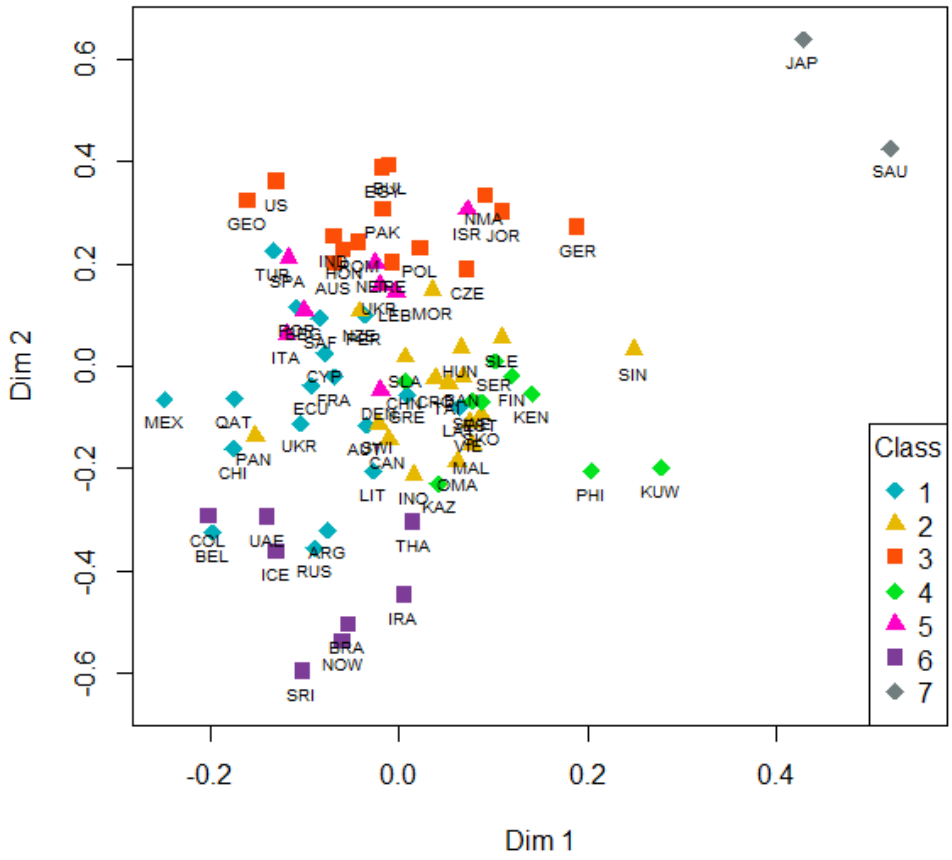


Figure 7: First two components of the factorial k-means analysis applied on the economic data set

this component. The second component is positively correlated with the percentage of children and adolescents, the crime rate, and the traffic commute time. The percentage of elderly people is negatively correlated with the second component. The third component is positively correlated with the quality of life and negatively correlated with the amount of pollution. The fourth component is only positively correlated and is explained by the health care quality and purchasing power.

Observing Figure 8 (page 18), we see that the first two components are plotted. When looking at the bottom right corner, we see that Iceland, Norway, Switzerland are clustered together. These countries are plotted to the right, because of their high cost of living and restaurant index, compared to the other countries. These variables are 100.48, 101.43 and 122.4 respectively for the cost of living index. The restaurant index equals 113.74, 109.28 and 123.01 respectively. Hong Kong has a relatively low cost of living and restaurant index. However, Hong Kong is plotted to the right because of the relatively large rent price. The rent index is 79.57, compared to the 46.95, 36.15 and 50.25 of Iceland, Norway and Switzerland respectively. The countries are plotted more to the

bottom because of their low crime rate and high percentage of elderly people. The crime index for Iceland, Norway, Switzerland and Hong Kong are 23.36, 35.43, 21.6 and 20.7 respectively. The percentage of elderly people are 14.4, 16.8, 18.4 and 16.3 respectively.

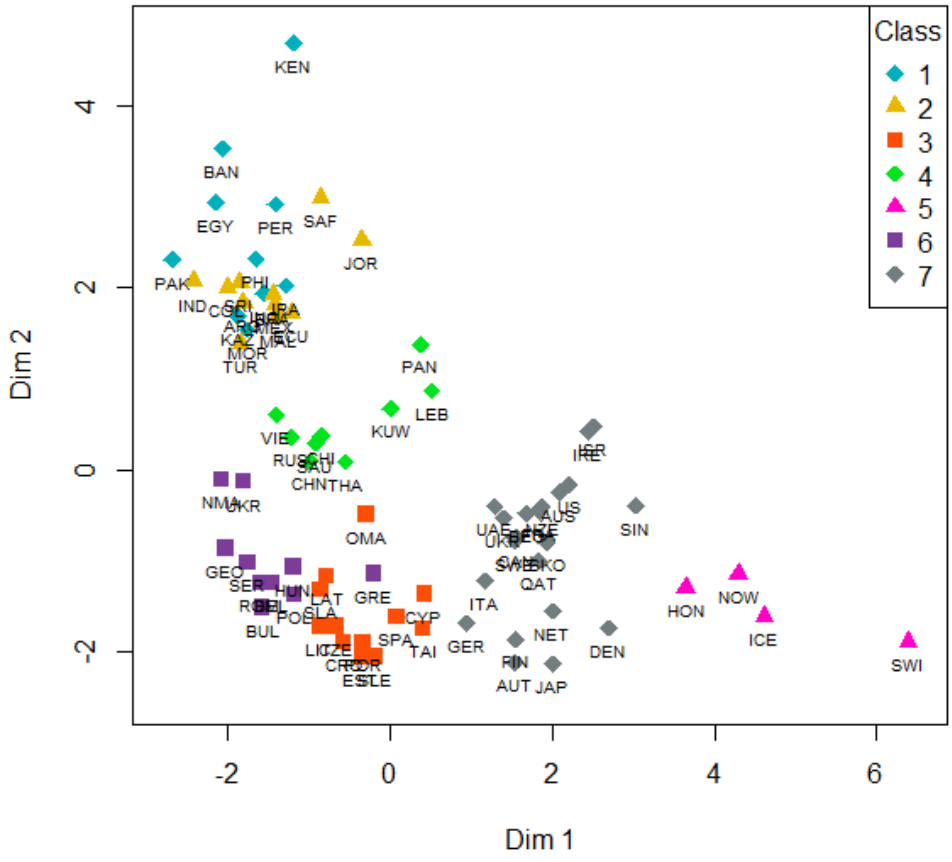


Figure 8: First two components of the CDPCA applied on the economic data set

When comparing the factorial k-means analysis and the CDPCA in terms of factor loadings (Table 7, page 16), the dimensions are not consistently explained through the same variables. A possible explaination for this phenomenon could be that the two methods ended up in different subspaces. When projecting the data set on a four dimensional subspace, the data points can be rotated in many different ways. On the basis of which initial values of $\mathbf{A}$, $\mathbf{U}$ and $\bar{\mathbf{Y}}$ (Section 2) are chosen, the algorithm can end up in different subspaces. This could explain the difference between the variables significantly describing different dimensions across the different methods.

# 5 Conclusion and Discussion

In this research, both clustering and dimension reduction analysis are compared. The research shows that the factorial k-means model outperforms the CDPCA and tandem analysis when applied on masked data. When both clustering and reducing the number of dimensions of the data, the factorial k-means best captures the relevant cluster structure. The CDPCA model performs effectively when handling masked data, but does not outperform the factorial k-means method. This can be explained due to the fact that the CDPCA analysis has more restrictions. The restriction makes it possible to also cluster the variables, and have each variable only explain one principal component.

The CDPCA being more restrictive, and giving worse results compared to the factorial k-means, would have one choose the factorial k-means over the CDPCA. However, the CDPCA restriction can be of great importance when trying to interpret the results, especially when handling large data sets with a large number of dimensions. Where the factors of the factorial k-means can be correlated to all the variables, the CDPCA has disjoint components.

The simulated data showed that the factorial k-means performs better, but the CDPCA does retrieve the original cluster structure and has an asset, in its ability to create disjoint components. To decide which method better to use, I compared the models using a economic data set. I know the CDPCA can make the interpretation of a large data set less complicated and this is shown in the results. Applying the CDPCA on the economic data gives a good clustering, which is easy to interpret because of the disjoint components.

When expanding on this research, there are a few things that could very well be expanded on. First, the number of clusters in this application are chosen following Vichi & Saporta (2009). Though the data sets are moderately similar in size, the optimal number of clusters could be chosen with more thought. This could be done by assessing the number of clusters based on the average silhouette width (Rousseeuw, 1987) or the Calinski-Harabasz (Caliński & Harabasz, 1974)).

Secondly, I could use a different method of choosing the initial values of $\mathbf{A}$, $\mathbf{U}$ and $\bar{\mathbf{Y}}$. In this research, I use random starting values for a large number of repetitions. Using a large number of random starting values decreases the probability of ending up in a local minimum. However, using smarter starting values might give better results. I could for example use standard k-means to retrieve a certain clustering, and use these as starting values.

In conclusion, for future studies, there are several propositions to further improve the methods discussed in this research. Nevertheless, by applying the factorial k-means analysis or the CDPCA several clustering problems that arose from the tandem approach have been eliminated. When putting factorial k-means or CDPCA into practice, the use of clustering and dimension reduction can be perfected.

## References


Arabie, P., & Hubert, L. (1994). Cluster analysis in marketing research. *Handbook of marketing research*.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, *3*(1), 1-27.

De Soete, G., & Carroll, J. D. (1994). K-means clustering in a low-dimensional euclidean space. *New Approaches in Classification and Data Analysis*, 212-219.

Gordon, A. (1999). *Classification, 2nd edition.* London: Chapman  Hall.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning.* New York: Springer.

Heij, C., de Boer, P., Franses, P. H., Kloek, T., & van Dijk, H. K. (2004). *Econometric methods with applications in business and economics.* Oxford: Oxford University Press.

Librero, A. B. N., Villardonand, P. G., & Freitas, A. (2019). *Bootstrap on classical biplots and clustering disjoint biplot.* (R reference manual)

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, *28*(2), 129-137.

Markos, A., D'Enza, A. I., & van de Velden, M. (2019). Beyond tandem analysis: Joint dimension reduction and clustering in r. *Journal of Statistical Software*, *91*(10).

Numbeo. (2020, June). *Cost of living.* `https://www.numbeo.com/cost-of-living/`. (Consulted on 08-06-2020)

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edingburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559-572.

Poole, D. (2006). *Linear algebra: A modern introduction.* Stamford: Cengage Learning.

Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65.

Vichi, M., & Kiers, H. A. (2001). Factorial k-means analysis for two-way data. *Computational Statistics & Data Analysis*, *37*(1), 49-64.

Vichi, M., & Saporta, G. (2009). Clustering and disjoint principal component analysis. *Computational Statistics & Data Analysis*, *53*(8), 3194-3208.

# Appendix A: R Codes Simulated Data

## Plot of first two variables

```
# This code plots the original two data variables and the equilateral
    triangle the centroids are located on

# This section makes the three originalClusters with a set distance and
    spread between the three centroids
distance = 6
spread = 1

originalCluster1 = read.csv("pointDistanceOLD.csv", header = FALSE)
originalCluster1 = originalCluster1*spread

originalCluster2 = originalCluster1
originalCluster2[,1] = originalCluster2[,1] + distance/2
originalCluster2[,2] = originalCluster2[,2] + (3^0.5)*(distance/2)

originalCluster3 = originalCluster1
originalCluster3[,1] = originalCluster3[,1] + distance

data <- rbind(rbind(originalCluster1, originalCluster2), originalCluster3)
data <- data[c(1,2,29,30,3:5,31:33,6:9,34:37,10:12,38:40,13,14,41,42,15:28
    ),]

# This section makes the maximum and minimum X and Y values for the plot
maxX = max(data[,1]) + distance/10
minX = min(data[,1]) - distance/10
maxY = max(data[,2]) + distance/10
minY = min(data[,2]) - distance/10

# The points and lines are plotted
plot(data, xlim=c(minX,maxX), ylim=c(minY,maxY), xlab="", ylab="", col= "
    black", pch=16, cex=1.4)
segments(0, 0, distance, 0, lty = 2, col="gray", lwd=2)
segments(0, 0, distance/2, (3^0.5)*(distance/2), lty = 2, col="gray", lwd=
    2)
segments(distance/2, (3^0.5)*(distance/2), distance, 0, lty = 2, col="gray
    ", lwd=2)
text(0, -2, labels="Cluster 1", cex= 0.9, pos=1)
text(distance, -2, labels="Cluster 3", cex= 0.9, pos=1)
text(3, (3^0.5)*(distance/2)-2, labels="Cluster 2", cex= 0.9, pos=1)
```

### K-means algorithm applied on the full simulated data set

```
# This code plots the k-means clusters of the data applied on the whole
    simulated data set

set.seed(27)

# This section makes the three originalClusters with a set distance and
    spread between the three centroids
distance = 6
spread = 1
cluster1 = read.csv("pointDistanceOLD.csv", header = FALSE)
cluster1 = cluster1*spread

cluster2 = cluster1
cluster2[,1] = cluster2[,1] + distance/2
cluster2[,2] = cluster2[,2] + (3^0.5)*(distance/2)

cluster3 = cluster1
cluster3[,1] = cluster3[,1] + distance

data <- rbind(rbind(cluster1, cluster2), cluster3)
data <- data[c(1,2,29,30,3:5,31:33,6:9,34:37,10:12,38:40,13,14,41,42,15:28
    ),]

# This section makes the maximum and minimum X and Y values for the plot
maxX = max(data[,1]) + distance/10
minX = min(data[,1]) - distance/10
maxY = max(data[,2]) + distance/10
minY = min(data[,2]) - distance/10

# I add the four normal drawn variables
data$V3 = c(rnorm(42, mean=0, sd=6))
data$V4 = c(rnorm(42, mean=0, sd=6))
data$V5 = c(rnorm(42, mean=0, sd=6))
data$V6 = c(rnorm(42, mean=0, sd=6))

# The points are plotted
plot = kmeans(data, 3)
plot(data[which(plot$cluster==1),c(1:2)], xlim=c(minX,maxX), ylim=c(minY,
    maxY), xlab="", ylab="", col= "#FC4E07", pch=15, cex=1.3)
points(data[which(plot$cluster==2),c(1:2)], xlab="", ylab="", col= "#E7B80
    0", pch=17, cex=1.3)
points(data[which(plot$cluster==3),c(1:2)], xlab="", ylab="", col= "#00
```

```
    AFBB", pch=18, cex=1.5)
legend("topleft", legend = c(1:3), col=c("#00AFBB", "#E7B800", "#FC4E07"),
       pch = c(18, 17, 15), pt.cex = c(1.5, 1.3, 1.3), cex=1, title ="Class")
text(0, -2, labels="Cluster 1", cex= 0.9, pos=1)
text(distance, -2, labels="Cluster 3", cex= 0.9, pos=1)
text(3, (3^0.5)*(distance/2)-2, labels="Cluster 2", cex= 0.9, pos=1)
```

### Tandem analysis applied on two-five principal components

```
# This code plots the four tandem analysis plots, with a different number
    of components
set.seed(27)

# This section makes the three originalClusters with a set distance and
    spread between the three centroids
distance = 6
spread = 1

cluster1 = read.csv("pointDistanceOLD.csv", header = FALSE)
cluster1 = cluster1*spread

cluster2 = cluster1
cluster2[,1] = cluster2[,1] + distance/2
cluster2[,2] = cluster2[,2] + (3^0.5)*(distance/2)

cluster3 = cluster1
cluster3[,1] = cluster3[,1] + distance

originalCluster = c(1,1,3,3,1,1,1,3,3,3,1,1,1,1,3,3,3,3,1,1,1,3,3,3,1,1,3,
    3,2,2,2,2,2,2,2,2,2,2,2,2,2,2)
data <- rbind(rbind(cluster1, cluster2), cluster3)
data <- data[c(1,2,29,30,3:5,31:33,6:9,34:37,10:12,38:40,13,14,41,42,15:28
    ),]

# This section makes the maximum and minimum X and Y values for the plot
maxX = max(data[,1]) + distance/10
minX = min(data[,1]) - distance/10
maxY = max(data[,2]) + distance/10
minY = min(data[,2]) - distance/10

# I add the four normal drawn variables
data$V3 = c(rnorm(42, mean=0, sd=6))
data$V4 = c(rnorm(42, mean=0, sd=6))
data$V5 = c(rnorm(42, mean=0, sd=6))
```

```r
data$V6 = c(rnorm(42, mean=0, sd=6))

# I apply PCA and create Table 1
pca = prcomp(data, center=TRUE, scale=TRUE)
table1 = matrix(1:24,6,4)
table1[,2] = (pca$sdev)^2
table1[,3] = table1[,2]/sum(table1[,2]) *100
table1[,4] = cumsum(table1[,2])/sum(table1[,2]) *100

# The points and clusters are plotted
par(mfrow=c(2,2))

plot2pca = kmeans(pca$x[,1:2], 3, iter.max=1000, nstart=1)
plot(pca$x[which(originalCluster==1),1:2], xlim=c(min(pca$x[,1])-0.5, max(
    pca$x[,1])+0.5), ylim=c(min(pca$x[,2])-0.5, max(pca$x[,2])+0.5), xlab="
    ", ylab="", col= "#E7B800", pch=17, cex=1.3, main="a", cex.main=1.7)
points(pca$x[which(originalCluster==2),1:2], xlim=c(minX,maxX), ylim=c(
    minY,maxY), xlab="", ylab="", col= "#FC4E07", pch=15, cex=1.3)
points(pca$x[which(originalCluster==3),1:2], xlim=c(minX,maxX), ylim=c(
    minY,maxY), xlab="", ylab="", col= "#00AFBB", pch=18, cex=1.5)

plot3pca = kmeans(pca$x[,1:3], 3, iter.max=1000, nstart=1)
plot(data[which(plot3pca$cluster==1),c(1:2)], xlim=c(minX,maxX), ylim=c(
    minY,maxY), xlab="", ylab="", col= "#E7B800", pch=17, cex=1.3, main="b"
    , cex.main=1.7)
points(data[which(plot3pca$cluster==2),c(1:2)], xlim=c(minX,maxX), ylim=c(
    minY,maxY), xlab="", ylab="", col= "#FC4E07", pch=15, cex=1.3)
points(data[which(plot3pca$cluster==3),c(1:2)], xlim=c(minX,maxX), ylim=c(
    minY,maxY), xlab="", ylab="", col= "#00AFBB", pch=18, cex=1.5)

plot4pca = kmeans(pca$x[,1:4], 3, iter.max=1000, nstart=1)
plot(data[which(plot4pca$cluster==1),c(1:2)], xlim=c(minX,maxX), ylim=c(
    minY,maxY), xlab="", ylab="", col= "#E7B800", pch=17, cex=1.3, main="c"
    , cex.main=1.7)
points(data[which(plot4pca$cluster==2),c(1:2)], xlim=c(minX,maxX), ylim=c(
    minY,maxY), xlab="", ylab="", col= "#00AFBB", pch=18, cex=1.5)
points(data[which(plot4pca$cluster==3),c(1:2)], xlim=c(minX,maxX), ylim=c(
    minY,maxY), xlab="", ylab="", col= "#FC4E07", pch=15, cex=1.3)

plot5pca = kmeans(pca$x[,1:5], 3, iter.max=1000, nstart=1)
plot(data[which(plot5pca$cluster==1),c(1:2)], xlim=c(minX,maxX), ylim=c(
    minY,maxY), xlab="", ylab="", col= "#E7B800", pch=17, cex=1.3, main="d"
    , cex.main=1.7)
```

```
points(data[which(plot5pca$cluster==2),c(1:2)], xlim=c(minX,maxX), ylim=c(minY,maxY), xlab="", ylab="", col= "#00AFBB", pch=18, cex=1.5)
points(data[which(plot5pca$cluster==3),c(1:2)], xlim=c(minX,maxX), ylim=c(minY,maxY), xlab="", ylab="", col= "#FC4E07", pch=15, cex=1.3)
```

### Clustering of the simulated data using factorial k-means

```
# This code plots the factorial k-means plot using simulated data

install.packages('clustrd')
library('clustrd')
set.seed(27)

# This section makes the three originalClusters with a set distance and spread between the three centroids
distance = 6
spread = 1

cluster1 = read.csv("pointDistanceOLD.csv", header = FALSE)
cluster1 = cluster1*spread

cluster2 = cluster1
cluster2[,1] = cluster2[,1] + distance/2
cluster2[,2] = cluster2[,2] + (3^0.5)*(distance/2)

cluster3 = cluster1
cluster3[,1] = cluster3[,1] + distance

originalCluster = c(1,1,3,3,1,1,1,3,3,3,1,1,1,1,3,3,3,3,1,1,1,3,3,3,1,1,3,3,2,2,2,2,2,2,2,2,2,2,2,2,2,2)
testData <- rbind(rbind(cluster1, cluster2), cluster3)
testData <- testData[c(1,2,29,30,3:5,31:33,6:9,34:37,10:12,38:40,13,14,41,42,15:28),]

# I add the four normal drawn variables
testData$V3 = c(rnorm(42, mean=0, sd=6))
testData$V4 = c(rnorm(42, mean=0, sd=6))
testData$V5 = c(rnorm(42, mean=0, sd=6))
testData$V6 = c(rnorm(42, mean=0, sd=6))

# I apply the factorial k-means method on the data with 3 clusters and 2 components.
outFKM = cluspca(testData, 3, 2, method="FKM", scale=F, center=F)

```

```
# The points and clusters are plotted
plot(outFKM$obscoord[which(originalCluster==1),c(1:2)], xlim=c(-2,8), ylim=c(-8,2), xlab="Dim 1", ylab="Dim 2", col= "#00AFBB", pch=18, cex=1.5)
points(outFKM$obscoord[which(originalCluster==2),c(1:2)], xlab="", ylab="", col= "#E7B800", pch=17, cex=1.3)
points(outFKM$obscoord[which(originalCluster==3),c(1:2)], xlab="", ylab="", col= "#FC4E07", pch=15, cex=1.3)
legend("bottomright", legend = c(1:3), col=c("#00AFBB", "#E7B800", "#FC4E07"), pch = c(18, 17, 15), pt.cex = c(1.5, 1.3, 1.3), title ="Class")

# I create table 2 with the factor loadings
table2 = t(outFKM$attcoord)
```

### Clustering of the simulated data using CDPCA

```
# This code plots the CDPCA plot using simulated data

install.packages('clustrd')
library('clustrd')
install.packages('biplotbootGUI')
library('biplotbootGUI')
set.seed(27)

# This section makes the three originalClusters with a set distance and spread between the three centroids
distance = 6
spread = 1

cluster1 = read.csv("pointDistanceOLD.csv", header = FALSE)
cluster1 = cluster1*spread

cluster2 = cluster1
cluster2[,1] = cluster2[,1] + distance/2
cluster2[,2] = cluster2[,2] + (3^0.5)*(distance/2)

cluster3 = cluster1
cluster3[,1] = cluster3[,1] + distance

originalCluster = c(1,1,3,3,1,1,1,3,3,3,1,1,1,1,3,3,3,3,1,1,1,3,3,3,1,1,3,3,2,2,2,2,2,2,2,2,2,2,2,2,2,2)
testData <- rbind(rbind(cluster1, cluster2), cluster3)
testData <- testData[c(1,2,29,30,3:5,31:33,6:9,34:37,10:12,38:40,13,14,41,42,15:28),]

```

```r
# I add the four normal drawn variables
testData$V3 = c(rnorm(42, mean=0, sd=6))
testData$V4 = c(rnorm(42, mean=0, sd=6))
testData$V5 = c(rnorm(42, mean=0, sd=6))
testData$V6 = c(rnorm(42, mean=0, sd=6))

# I apply the CDPCA method on the data with 3 clusters and 2 components.
outFKM = CDpca(testData, P=3, Q=2, maxit=100, r=30, cdpcaplot=F)

# The points and clusters are plotted
plot(outFKM$Y[which(originalCluster==1),c(1:2)], xlim=c(-2,2.2), ylim=c(-2
    ,2.4), xlab="", ylab="", col= "#00AFBB", pch=18, cex=1.5)
points(outFKM$Y[which(originalCluster==2),c(1:2)], xlab="", ylab="", col=
    "#E7B800", pch=17, cex=1.3)
points(outFKM$Y[which(originalCluster==3),c(1:2)], xlab="", ylab="", col=
    "#FC4E07", pch=15, cex=1.3)
legend("bottomright", legend = c(1:3), col=c("#00AFBB", "#E7B800", "#FC4E0
    7"), pch = c(18, 17, 15), pt.cex = c(1.5, 1.3, 1.3), title ="Class")

# I create table 3 with the factor loadings
table3 = t(outFKM$A)
```

## Appendix B: Simulated Data

| Original | | Mask | | | |
|---|---|---|---|---|---|
| V1 | V2 | V3 | V4 | V5 | V6 |
| -0.5 | -1.739 | -3.630 | 3.292 | 1.257 | 14.779 |
| 0.5 | -1.739 | 4.991 | 0.242 | 9.713 | -0.486 |
| 5.5 | -1.739 | -5.280 | -7.262 | -0.193 | 4.156 |
| 6.5 | -1.739 | 3.821 | 3.639 | -0.874 | 8.712 |
| -1 | -0.872 | -6.399 | 7.515 | -2.472 | -1.475 |
| 0 | -0.872 | 3.503 | -6.861 | 11.383 | -0.083 |
| 1 | -0.872 | 2.600 | -0.175 | -1.302 | 4.005 |
| 5 | -0.872 | 6.930 | 7.681 | 12.996 | -5.981 |
| 6 | -0.872 | 6.575 | -2.603 | -10.026 | 3.838 |
| 7 | -0.872 | 2.048 | 9.850 | 0.109 | -0.769 |
| -1.5 | 0.001 | 11.637 | 6.176 | 3.715 | -0.888 |
| -0.5 | 0.001 | -0.038 | 5.629 | -1.024 | 0.220 |
| 0.5 | 0.001 | 5.296 | 3.611 | 0.407 | -0.298 |
| 1.5 | 0.001 | 2.819 | 10.362 | 10.511 | 5.849 |
| 4.5 | 0.001 | -8.798 | 6.907 | -12.112 | -6.806 |
| 5.5 | 0.001 | -5.933 | 4.020 | -1.496 | -2.932 |
| 6.5 | 0.001 | -7.377 | 4.077 | 9.395 | 1.467 |
| 7.5 | 0.001 | 0.968 | -2.164 | 3.183 | 6.046 |
| -1 | 0.868 | 4.861 | 6.979 | -5.299 | -9.921 |
| 0 | 0.868 | 1.585 | -1.985 | -3.296 | -0.794 |
| 1 | 0.868 | -7.757 | 11.958 | -4.476 | -0.951 |
| 5 | 0.868 | 2.634 | -0.889 | 3.000 | -2.885 |
| 6 | 0.868 | -3.436 | -4.900 | -0.658 | 1.326 |
| 7 | 0.868 | -2.237 | -4.301 | -2.810 | -4.241 |
| -0.5 | 1.741 | -6.860 | -1.038 | 4.055 | -12.914 |
| 0.5 | 1.741 | -3.001 | 8.130 | 5.793 | -7.786 |
| 5.5 | 1.741 | -2.320 | 5.710 | 1.652 | 0.280 |
| 6.5 | 1.741 | -4.227 | -1.768 | -2.696 | 2.933 |
| 2.5 | 3.458 | 8.260 | 10.803 | 4.362 | -1.990 |
| 3.5 | 3.458 | -7.696 | 0.914 | 7.987 | -7.406 |
| 2 | 4.324 | -7.608 | -5.149 | 0.558 | -0.452 |
| 3 | 4.324 | 0.903 | 2.618 | -5.365 | 6.603 |
| 4 | 4.324 | 1.409 | -8.405 | -5.876 | -1.173 |
| 1.5 | 5.198 | 1.317 | 1.740 | -1.232 | -0.319 |

**Table continued from previous page**

| Original | | Mask | | | |
|---|---|---|---|---|---|
| V1 | V2 | V3 | V4 | V5 | V6 |
| 2.5 | 5.198 | -2.324 | 4.840 | 0.339 | 14.030 |
| 3.5 | 5.198 | 5.508 | -5.433 | 6.023 | 3.027 |
| 4.5 | 5.198 | -15.216 | 0.448 | -6.292 | 2.401 |
| 2 | 6.064 | 0.919 | 2.186 | -1.074 | -11.174 |
| 3 | 6.064 | -8.402 | -5.312 | -6.717 | 12.337 |
| 4 | 6.064 | -1.664 | -16.231 | -5.689 | 4.523 |
| 2.5 | 6.938 | -2.500 | -0.128 | -3.445 | -10.903 |
| 3.5 | 6.938 | -5.279 | -3.637 | 5.062 | -4.722 |

## Appendix C: R Codes Economic Data

### Tandem analysis applied on the economic data set

```
# This code plots the tandem plot using economic data

install.packages('clustrd')
library('clustrd')
set.seed(2101)

# I load in the data
originalData = read.csv("DATASET.csv", header = TRUE)
data = originalData[,c(3:12,14:15)]

# Specify the number of clusters and components
NoC = 7
NoD = 4

# Apply the pca and kmeans on the economic data
pca = prcomp(data, center=T, scale=TRUE)
plot2pca = kmeans(pca$x[,1:NoD], NoC, iter.max=1000, nstart=1)
plot2 = data.frame(c(plot2pca$cluster))

# The points and clusters are plotted
plot(pca$x[which(plot2==1),1:2],
     xlim=c(-5,6.7), ylim=c(-2.5,2.8),
     xlab="Dim 1", ylab="Dim 2", col= "#00AFBB", pch=18, cex=1.5)
points(pca$x[which(plot2==2),1:2], xlab="", ylab="", col= "#E7B800", pch=1
    7, cex=1.3)
points(pca$x[which(plot2==3),1:2], xlab="", ylab="", col= "#FC4E07", pch=1
    5, cex=1.3)
points(pca$x[which(plot2==4),1:2], xlab="", ylab="", col= "#00E31E", pch=1
    8, cex=1.5)
points(pca$x[which(plot2==5),1:2], xlab="", ylab="", col= "#FF00C6", pch=1
    7, cex=1.3)
points(pca$x[which(plot2==6),1:2], xlab="", ylab="", col= "#7D3C98", pch=1
    5, cex=1.3)
points(pca$x[which(plot2==7),1:2], xlab="", ylab="", col= "#717D7E", pch=1
    8, cex=1.5)
legend("bottomright", legend = c(1:7), col=c("#00AFBB", "#E7B800", "#FC4E0
    7", "#00E31E", "#FF00C6", "#7D3C98", "#717D7E"), pch = c(18, 17, 15, 18
    , 17, 15, 18), pt.cex = c(1.5, 1.3, 1.3, 1.5, 1.3, 1.3, 1.5), title ="
    Class")
text(pca$x[,1], pca$x[,2], labels=originalData[,2], cex= 0.6, pos=1)
```

```
32
33 # I create table 7a with the factor loadings
34 table7a = pca$rotation[,1:NoD]
```

### Factorial k-means analysis applied on the economic data set

```
1  # This code plots the factorial k-means plot using economic data
2
3  install.packages('clustrd')
4  library('clustrd')
5  set.seed(2101)
6
7  # I load in the data
8  originalData = read.csv("DATASET.csv", header = TRUE)
9  data = originalData[,c(3:12,14:15)]
10
11 # Specify the number of clusters and components
12 NoC = 7
13 NoD = 4
14
15 # Apply the factorial k-means on the economic data
16 outFKM = cluspca(data, NoC, NoD, method="FKM", scale=T, center=T, alpha=0,
       nstart=100)
17 plot = data.frame(c(outFKM$cluster))
18
19 # The points and clusters are plotted
20 plot(outFKM$obscoord[which(plot==1),1:2], main="",
21      xlim=c(-0.25,0.55), ylim=c(-0.65,0.65),
22      #xlim=c(-1,1.2), ylim=c(-0.35,0.3),
23      xlab="Dim 1", ylab="Dim 2", col= "#00AFBB", pch=18, cex=1.5)
24 points(outFKM$obscoord[which(plot==2),1:2], xlab="", ylab="", col= "#E7B80
     0", pch=17, cex=1.3)
25 points(outFKM$obscoord[which(plot==3),1:2], xlab="", ylab="", col= "#FC4E0
     7", pch=15, cex=1.3)
26 points(outFKM$obscoord[which(plot==4),1:2], xlab="", ylab="", col= "#00E31
     E", pch=18, cex=1.5)
27 points(outFKM$obscoord[which(plot==5),1:2], xlab="", ylab="", col= "#FF00C
     6", pch=17, cex=1.3)
28 points(outFKM$obscoord[which(plot==6),1:2], xlab="", ylab="", col= "#7D3C9
     8", pch=15, cex=1.3)
29 points(outFKM$obscoord[which(plot==7),1:2], xlab="", ylab="", col= "#717D7
     E", pch=18, cex=1.5)
30 legend("bottomright", legend = c(1:7), col=c("#00AFBB", "#E7B800", "#FC4E0
     7", "#00E31E", "#FF00C6", "#7D3C98", "#717D7E"), pch = c(18, 17, 15, 18
```

```
    , 17, 15, 18), pt.cex = c(1.5, 1.3, 1.3, 1.5, 1.3, 1.3, 1.5), title ="
    Class")
text(outFKM$obscoord[,1], outFKM$obscoord[,2], labels=originalData[,2],
    cex= 0.6, pos=1)

# I create table 7b with the factor loadings
tableFACK = t(outFKM$attcoord)
```

### CDPCA applied on the economic data set

```
# This code plots the CDPCA plot using economic data

install.packages('clustrd')
install.packages('biplotbootGUI')
library('biplotbootGUI')
library('clustrd')
set.seed(2101)

# I load in the data
data  = read.csv("DATASET.csv", header = TRUE)

# Specify the number of clusters and components
NoC = 7
NoD = 4

# Apply the CDPCA on the economic data
outCDPCA = CDpca(data[,c(3:12,14:15)], P=NoC, Q=NoD, maxit=1000, r=30,
    cdpcaplot=F)
plot = data.frame(c(outCDPCA$tableclass))

# The points and clusters are plotted
plot(outCDPCA$Y[which(plot==1),1:2],
     xlim=c(-2.8,6.5), ylim=c(-2.5,4.8),
     xlab="Dim 1", ylab="Dim 2", col= "#00AFBB", pch=18, cex=1.5)
points(outCDPCA$Y[which(plot==2),1:2], xlab="", ylab="", col= "#E7B800",
    pch=17, cex=1.3)
points(outCDPCA$Y[which(plot==3),1:2], xlab="", ylab="", col= "#FC4E07",
    pch=15, cex=1.3)
points(outCDPCA$Y[which(plot==4),1:2], xlab="", ylab="", col= "#00E31E",
    pch=18, cex=1.5)
points(outCDPCA$Y[which(plot==5),1:2], xlab="", ylab="", col= "#FF00C6",
    pch=17, cex=1.3)
points(outCDPCA$Y[which(plot==6),1:2], xlab="", ylab="", col= "#7D3C98",
    pch=15, cex=1.3)
```

```
29 points(outCDPCA$Y[which(plot==7),1:2], xlab="", ylab="", col= "#717D7E",
       pch=18, cex=1.5)
30 legend("topright", legend = c(1:7), col=c("#00AFBB", "#E7B800", "#FC4E07",
        "#00E31E", "#FF00C6", "#7D3C98", "#717D7E"), pch = c(18, 17, 15, 18, 1
       7, 15, 18), pt.cex = c(1.5, 1.3, 1.3, 1.5, 1.3, 1.3, 1.5), title ="
       Class")
31 text(outCDPCA$Y[,1], outCDPCA$Y[,2], labels=data[,2], cex= 0.6, pos=1)
32
33 # I create table 7c with the factor loadings
34 tableCDPCA = t(outCDPCA$A)
```

# Appendix D: Economic Data

### First six variables

| | Age 0 to 14 Years | Age above 65 Years | Crime Index | Cost of Living Index | Rent Index | Groceries Index |
|---|---|---|---|---|---|---|
| ARG | 24.90 | 11.20 | 61.77 | 32.95 | 8.33 | 25.82 |
| AUS | 19.00 | 15.50 | 41.36 | 73.54 | 34.86 | 67.23 |
| AUT | 14.10 | 19.20 | 23.73 | 70.38 | 26.81 | 61.73 |
| BAN | 28.40 | 5.10 | 63.94 | 32.25 | 5.01 | 29.72 |
| BEL | 16.70 | 14.80 | 24.99 | 34.70 | 10.50 | 28.06 |
| BEG | 17.10 | 18.60 | 43.98 | 71.78 | 25.43 | 58.66 |
| BRA | 21.70 | 8.60 | 68.88 | 40.22 | 10.65 | 29.20 |
| BUL | 14.20 | 20.80 | 38.50 | 36.70 | 9.64 | 30.09 |
| CAN | 16.00 | 17.00 | 39.67 | 67.62 | 30.73 | 63.68 |
| CHI | 20.30 | 11.10 | 45.23 | 43.62 | 13.39 | 36.45 |
| CHN | 17.70 | 10.60 | 31.83 | 40.04 | 16.38 | 40.37 |
| COL | 23.50 | 7.60 | 54.79 | 30.66 | 9.58 | 25.05 |
| CRO | 14.70 | 19.70 | 24.71 | 49.70 | 13.50 | 39.90 |
| CYP | 16.80 | 13.40 | 30.01 | 57.93 | 20.54 | 44.21 |
| CZE | 15.40 | 19.00 | 25.52 | 46.15 | 19.56 | 38.13 |
| DEN | 16.50 | 19.70 | 25.10 | 83.00 | 31.92 | 61.74 |
| ECU | 28.40 | 7.10 | 50.90 | 40.98 | 11.96 | 35.46 |
| EGY | 33.50 | 5.10 | 46.92 | 29.54 | 5.49 | 25.50 |
| EST | 16.40 | 19.40 | 23.14 | 50.93 | 15.41 | 36.57 |
| FIN | 16.40 | 21.20 | 23.32 | 70.29 | 26.16 | 56.52 |
| FRA | 18.10 | 19.70 | 46.79 | 74.14 | 25.39 | 67.90 |
| GEO | 19.20 | 14.80 | 20.21 | 28.48 | 9.80 | 23.05 |
| GER | 13.10 | 21.40 | 34.81 | 65.26 | 27.06 | 49.23 |
| GRE | 14.20 | 20.40 | 40.32 | 55.67 | 11.68 | 41.63 |
| HON | 11.50 | 16.30 | 20.70 | 77.22 | 79.57 | 75.94 |
| HUN | 14.30 | 18.60 | 35.08 | 40.85 | 13.97 | 30.77 |
| ICE | 20.10 | 14.40 | 23.36 | 100.48 | 46.95 | 86.89 |
| IND | 27.80 | 6.00 | 43.32 | 24.58 | 5.68 | 24.55 |
| INO | 27.40 | 5.30 | 45.84 | 37.27 | 10.62 | 37.36 |
| IRA | 23.70 | 5.40 | 49.25 | 39.01 | 14.48 | 35.54 |
| IRE | 21.60 | 14.00 | 45.43 | 75.91 | 43.88 | 58.35 |
| ISR | 27.90 | 11.70 | 29.60 | 81.15 | 31.33 | 66.31 |
| ITA | 13.50 | 23.00 | 44.26 | 67.26 | 21.22 | 55.44 |

**Table continued from previous page**

| | Age 0 to 14 Years | Age above 65 Years | Crime Index | Cost of Living Index | Rent Index | Groceries Index |
|---|---|---|---|---|---|---|
| JAP | 12.90 | 27.00 | 20.66 | 83.35 | 25.97 | 81.82 |
| JOR | 35.50 | 3.80 | 40.83 | 53.67 | 11.54 | 43.32 |
| KAZ | 27.90 | 7.00 | 62.02 | 30.64 | 9.78 | 24.31 |
| KEN | 40.50 | 2.70 | 61.66 | 40.21 | 10.73 | 35.35 |
| KUW | 21.10 | 2.30 | 34.75 | 50.37 | 31.21 | 34.68 |
| LAT | 15.40 | 19.80 | 36.95 | 47.94 | 12.34 | 34.85 |
| LEB | 23.10 | 8.50 | 43.36 | 60.50 | 24.54 | 43.92 |
| LIT | 14.80 | 19.00 | 33.06 | 44.28 | 13.65 | 33.63 |
| MAL | 24.30 | 6.30 | 58.84 | 39.12 | 11.00 | 37.58 |
| MEX | 26.70 | 6.80 | 53.97 | 35.72 | 11.46 | 32.39 |
| MOR | 27.40 | 6.80 | 48.69 | 34.32 | 8.94 | 30.11 |
| NET | 16.40 | 18.80 | 27.62 | 73.75 | 35.18 | 55.87 |
| NZE | 19.80 | 15.30 | 40.93 | 72.53 | 32.09 | 64.69 |
| NMA | 23.70 | 11.90 | 38.67 | 31.59 | 6.49 | 24.94 |
| NOW | 17.80 | 16.80 | 35.43 | 101.43 | 36.15 | 91.14 |
| OMA | 21.80 | 2.40 | 20.79 | 49.28 | 17.98 | 43.50 |
| PAK | 34.80 | 4.50 | 44.08 | 21.98 | 4.59 | 19.08 |
| PAN | 27.40 | 7.90 | 47.19 | 54.16 | 24.76 | 53.03 |
| PER | 27.40 | 7.10 | 68.15 | 38.65 | 12.78 | 33.80 |
| PHI | 31.70 | 4.80 | 42.16 | 37.63 | 9.00 | 33.46 |
| POL | 14.80 | 16.80 | 28.50 | 40.04 | 15.67 | 30.55 |
| POR | 13.60 | 21.50 | 29.63 | 49.52 | 21.81 | 38.14 |
| QAT | 13.90 | 1.30 | 11.86 | 64.04 | 47.44 | 53.61 |
| ROM | 15.30 | 17.80 | 27.64 | 35.31 | 10.05 | 29.03 |
| RUS | 17.60 | 14.20 | 41.12 | 39.21 | 11.36 | 31.08 |
| SAU | 25.20 | 3.30 | 26.18 | 48.34 | 11.39 | 37.89 |
| SER | 16.50 | 17.30 | 37.41 | 35.72 | 9.08 | 25.46 |
| SIN | 15.00 | 12.90 | 30.57 | 81.10 | 63.27 | 66.75 |
| SLA | 15.40 | 15.00 | 29.22 | 44.46 | 16.11 | 37.51 |
| SLE | 15.00 | 19.00 | 21.07 | 53.43 | 17.09 | 43.76 |
| SAF | 29.00 | 5.30 | 77.49 | 42.87 | 16.61 | 33.29 |
| SKO | 13.50 | 13.90 | 28.02 | 78.18 | 22.86 | 91.31 |
| SPA | 14.70 | 19.40 | 31.96 | 53.77 | 21.77 | 42.38 |
| SRI | 24.00 | 10.10 | 40.22 | 31.61 | 7.74 | 35.09 |
| SWE | 17.50 | 20.00 | 47.07 | 69.85 | 25.90 | 60.47 |

**Table continued from previous page**

| | Age 0 to 14 Years | Age above 65 Years | Crime Index | Cost of Living Index | Rent Index | Groceries Index |
|---|---|---|---|---|---|---|
| SWI | 14.90 | 18.40 | 21.60 | 122.40 | 50.25 | 120.27 |
| TAI | 13.12 | 13.86 | 15.65 | 61.37 | 16.42 | 71.51 |
| THA | 17.30 | 11.40 | 40.48 | 49.77 | 17.10 | 49.20 |
| TUR | 25.00 | 8.10 | 39.49 | 34.69 | 6.78 | 29.76 |
| UKR | 15.50 | 16.50 | 48.85 | 33.18 | 10.46 | 26.01 |
| UAE | 13.90 | 1.10 | 15.70 | 61.98 | 41.07 | 47.63 |
| UK | 17.70 | 18.50 | 43.71 | 67.28 | 29.85 | 51.27 |
| US | 18.90 | 15.40 | 47.20 | 71.05 | 40.32 | 66.61 |
| VIE | 23.10 | 7.10 | 45.35 | 38.34 | 13.57 | 37.06 |

### Last six variables

| | Restaurant Price Index | Health Care Index | Quality of Life Index | Purchasing Power Index | Traffic Commute Time Index | Pollution Index |
|---|---|---|---|---|---|---|
| ARG | 30.52 | 69.25 | 115.31 | 47.22 | 43.08 | 50.67 |
| AUS | 70.32 | 77.38 | 186.21 | 107.31 | 34.73 | 23.46 |
| AUT | 68.09 | 78.73 | 182.50 | 82.38 | 26.27 | 22.19 |
| BAN | 19.21 | 42.80 | 70.03 | 33.21 | 56.73 | 86.21 |
| BEL | 38.94 | 59.04 | 134.83 | 37.00 | 30.68 | 43.63 |
| BEG | 80.14 | 74.34 | 153.47 | 86.28 | 36.15 | 52.94 |
| BRA | 30.78 | 56.29 | 105.65 | 32.81 | 41.70 | 54.98 |
| BUL | 29.14 | 55.40 | 129.80 | 49.37 | 29.39 | 65.33 |
| CAN | 63.53 | 71.58 | 163.47 | 95.09 | 33.87 | 27.83 |
| CHI | 41.57 | 65.44 | 119.76 | 42.50 | 35.44 | 65.78 |
| CHN | 29.16 | 64.48 | 102.81 | 60.88 | 41.81 | 80.77 |
| COL | 22.44 | 67.24 | 105.83 | 31.12 | 47.49 | 62.83 |
| CRO | 42.44 | 62.68 | 159.01 | 50.42 | 29.11 | 30.46 |
| CYP | 63.11 | 51.75 | 147.93 | 57.41 | 23.95 | 53.55 |
| CZE | 34.10 | 74.62 | 156.24 | 62.82 | 29.65 | 40.23 |
| DEN | 100.75 | 80.00 | 192.67 | 100.88 | 28.85 | 21.33 |
| ECU | 30.66 | 70.59 | 125.14 | 36.08 | 37.55 | 57.00 |
| EGY | 23.53 | 45.84 | 86.54 | 22.41 | 49.78 | 85.65 |
| EST | 52.94 | 72.67 | 177.82 | 71.30 | 24.53 | 19.81 |
| FIN | 76.81 | 75.79 | 190.22 | 99.93 | 29.90 | 11.55 |
| FRA | 72.54 | 79.99 | 153.95 | 80.36 | 34.76 | 43.56 |

**Table continued from previous page**

| | Restaurant Price Index | Health Care Index | Quality of Life Index | Purchasing Power Index | Traffic Commute Time Index | Pollution Index |
|---|---|---|---|---|---|---|
| GEO | 26.66 | 51.24 | 115.95 | 24.88 | 36.02 | 71.09 |
| GER | 61.58 | 73.32 | 179.78 | 102.36 | 30.98 | 29.03 |
| GRE | 53.74 | 56.21 | 133.07 | 43.68 | 33.84 | 52.55 |
| HON | 54.36 | 66.08 | 99.05 | 65.32 | 41.46 | 67.69 |
| HUN | 34.46 | 47.80 | 128.16 | 47.55 | 35.78 | 48.29 |
| ICE | 113.74 | 65.92 | 181.75 | 79.44 | 20.10 | 16.21 |
| IND | 17.17 | 67.13 | 108.63 | 54.30 | 46.99 | 78.87 |
| INO | 18.25 | 60.48 | 97.47 | 25.05 | 43.11 | 66.56 |
| IRA | 25.12 | 51.70 | 74.14 | 22.69 | 48.01 | 77.45 |
| IRE | 81.24 | 51.89 | 153.53 | 80.88 | 37.68 | 33.99 |
| ISR | 88.65 | 73.29 | 149.94 | 78.09 | 35.91 | 57.25 |
| ITA | 72.32 | 66.59 | 140.76 | 65.59 | 34.42 | 55.63 |
| JAP | 48.95 | 81.14 | 167.99 | 87.28 | 39.15 | 39.59 |
| JOR | 47.12 | 64.60 | 112.40 | 34.88 | 42.03 | 77.78 |
| KAZ | 29.87 | 50.70 | 88.31 | 38.36 | 31.08 | 75.15 |
| KEN | 35.06 | 55.59 | 70.56 | 27.32 | 56.65 | 76.60 |
| KUW | 47.08 | 56.21 | 115.75 | 85.59 | 34.44 | 68.69 |
| LAT | 42.88 | 62.91 | 150.00 | 52.48 | 32.89 | 33.73 |
| LEB | 58.16 | 64.38 | 106.56 | 44.47 | 37.38 | 88.37 |
| LIT | 42.70 | 69.49 | 159.42 | 57.85 | 26.08 | 28.80 |
| MAL | 22.94 | 68.10 | 118.44 | 64.49 | 37.03 | 63.18 |
| MEX | 32.12 | 70.12 | 118.55 | 41.81 | 39.39 | 66.10 |
| MOR | 24.50 | 45.72 | 105.46 | 35.25 | 36.89 | 70.64 |
| NET | 80.48 | 74.65 | 183.67 | 90.73 | 29.43 | 27.41 |
| NZE | 68.80 | 73.81 | 181.02 | 92.66 | 31.10 | 23.40 |
| NMA | 23.31 | 56.38 | 110.46 | 37.15 | 27.61 | 80.23 |
| NOW | 109.28 | 74.36 | 175.19 | 88.38 | 26.99 | 20.35 |
| OMA | 44.26 | 58.15 | 167.09 | 80.97 | 22.80 | 37.74 |
| PAK | 16.78 | 60.59 | 105.44 | 30.57 | 38.56 | 74.25 |
| PAN | 47.27 | 59.93 | 108.36 | 34.23 | 36.48 | 63.09 |
| PER | 24.69 | 56.15 | 85.46 | 33.95 | 48.33 | 84.13 |
| PHI | 20.84 | 67.47 | 85.37 | 23.48 | 44.63 | 74.28 |
| POL | 33.45 | 61.01 | 141.83 | 59.61 | 31.72 | 54.46 |
| POR | 42.48 | 71.88 | 162.91 | 49.43 | 30.00 | 30.89 |

**Table continued from previous page**

| | Restaurant Price Index | Health Care Index | Quality of Life Index | Purchasing Power Index | Traffic Commute Time Index | Pollution Index |
|---|---|---|---|---|---|---|
| QAT | 66.83 | 73.30 | 162.29 | 111.69 | 29.72 | 61.06 |
| ROM | 30.50 | 55.06 | 132.44 | 48.86 | 34.75 | 58.42 |
| RUS | 39.61 | 57.59 | 102.31 | 38.94 | 45.30 | 62.79 |
| SAU | 33.51 | 59.11 | 150.56 | 100.00 | 28.61 | 65.09 |
| SER | 28.63 | 51.27 | 116.30 | 36.72 | 30.43 | 60.32 |
| SIN | 58.99 | 70.84 | 144.39 | 88.96 | 41.31 | 33.48 |
| SLA | 33.82 | 60.02 | 152.53 | 56.94 | 29.11 | 39.66 |
| SLE | 45.43 | 64.58 | 172.15 | 66.31 | 27.30 | 24.06 |
| SAF | 40.58 | 64.14 | 131.97 | 73.61 | 39.43 | 57.30 |
| SKO | 44.87 | 81.97 | 139.02 | 85.21 | 39.57 | 62.48 |
| SPA | 52.07 | 78.88 | 169.82 | 72.03 | 29.10 | 39.99 |
| SRI | 18.76 | 72.53 | 85.16 | 24.45 | 59.01 | 59.14 |
| SWE | 72.35 | 69.23 | 175.95 | 101.73 | 30.49 | 18.09 |
| SWI | 123.01 | 72.44 | 192.01 | 119.53 | 29.09 | 22.39 |
| TAI | 28.49 | 86.71 | 143.23 | 65.67 | 31.93 | 63.35 |
| THA | 24.90 | 77.95 | 101.88 | 35.45 | 38.23 | 75.07 |
| TUR | 24.16 | 69.80 | 127.10 | 40.85 | 44.65 | 67.35 |
| UKR | 26.42 | 52.33 | 104.77 | 31.80 | 38.65 | 65.08 |
| UAE | 61.32 | 67.04 | 156.67 | 91.58 | 36.85 | 51.15 |
| UK | 74.13 | 74.46 | 162.71 | 91.73 | 34.53 | 40.56 |
| US | 70.74 | 69.27 | 172.11 | 109.52 | 32.89 | 36.88 |
| VIE | 19.59 | 57.70 | 87.48 | 28.14 | 30.17 | 86.47 |